EXHIBIT 3

CEBM

CEBM UNIVERSITY OF OXFORD

Explanation of the 2011 OCEBM Levels of Evidence

2 11

Introduction

The OCEBM Levels of Evidence was designed so that in addition to traditional critical appraisal, it can be used as a heuristic that clinicians and patients can use to answer clinical questions quickly and without resorting to pre-appraised sources. Heuristics are essentially rules of thumb that helps us make a decision in real environments, and are often as accurate as a more complicated decision process.

A distinguishing feature is that the Levels cover the entire range of clinical questions, in the order (from top row to bottom row) that the clinician requires. While most ranking schemes consider strength of evidence for therapeutic effects and harms, the OCEBM system allows clinicians and patients to appraise evidence for prevalence, accuracy of diagnostic tests, prognosis, therapeutic effects, rare harms, common harms, and usefulness of (early) screening.

Pre-appraised sources such as the Trip Database (1), (2), or REHAB+ (3) are useful because people who have the time and expertise to conduct systematic reviews of all the evidence design them. At the same time, clinicians, patients, and others may wish to keep some of the power over critical appraisal in their own hands.

History

Evidence ranking schemes have been used, and criticised, for decades (4-7), and each scheme is geared to answer different questions(8). Early evidence hierarchies(5, 6, 9) were introduced primarily to help clinicians and other researchers appraise the quality of evidence for therapeutic effects, while more recent attempts to assign levels to evidence have been designed to help systematic reviewers(8), or guideline developers(10).

While they are simple and easy to use, early hierarchies that placed randomized trials categorically above observational studies were criticized(11) for being simplistic(12). In some cases, observational studies give us the 'best' evidence(11). For example, there is a growing recognition that observational studies – even case-series (13) and *anecdotes* can sometimes provide definitive evidence(14).

More recent evidence-ranking schemes such as GRADE avoid this common objection by allowing observational studies with dramatic effects to be 'upgraded' (12), and trials may be 'downgraded' for quality and other reasons. Another advantage of the GRADE approach is that it takes other important factors such as directness, precision, and consistency when appraising quality of evidence. However, what GRADE has gained in accuracy, it may have lost in simplicity and efficiency. The GRADE system takes time to master and moreover is intended for appraising systematic reviews used in the production of guidelines.

VIDEO: CEBM 20 YEARS ON

CEBM 20 years on.

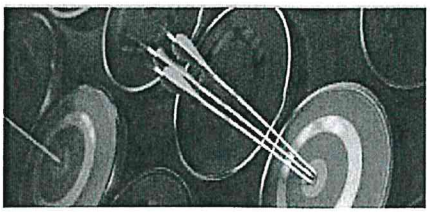
Carl Heneghan on two decades of promoting the practice and teaching of EBM.

RECOMMENDED CONTENT

OCEBM Levels of Evidence

The Levels of Evidence help you to target your search at the type of evidence that is most likely to provide a reliable answer.

Levels of Evidence: Introductory Document



Meanwhile, busy clinicians, who have only have a few minutes to answer a clinical question, will need a "fast and frugal" heuristic search tool to find and use the likely best evidence (15, 16).

The original CEBM Levels was first released in September 2000 for *Evidence-Based On Call* to make the process of finding evidence feasible and its results explicit. Busy clinicians sometimes need to make decisions quickly, sometimes in the middle of the night. One problem with many of the evidence ranking schemes at the time was that they ranked evidence for therapy and prevention, but not evidence for diagnostic tests, prognostic markers, or harm. A team led by Bob Phillips and Chris Ball, which included Dave Sackett, Doug Badenoch, Sharon Straus, Brian Haynes, and Martin Dawes therefore produced a Levels that included levels of evidence for therapy/prevention/aetiology/harm, prognosis, diagnosis, differential diagnosis, and economic and decision analyses.

While still useful as they are, in 2009 the Levels was over a decade old, and feedback over the years about the Levels led members of the OCEBM to believe it was time to review them. An international team led by Jeremy Howick (with considerable help from Olive Goddard and Mary Hodgkinson) that included Iain Chalmers, Paul Glasziou (chair), Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, and Hazel Thornton met for 2 days in Oxford in December 2009 to discuss potential changes to the OCEBM Levels.

After brainstorming for a few hours, the group voted on what they thought required revision. The following emerged as essential to developing a revised Evidence Levels:

1. That the Levels should be designed in a way that they could be used as a *search heuristic* for busy clinicians and patients to use in real time in addition to serving as a hierarchy of evidence. With that in mind, they simplified the Levels in several ways. For example, levels '1a', '1b', and '1c', in the original Levels was replaced with simply '1'. It was also modified to represent the natural flow of a clinical encounter (diagnosis, prognosis, treatment, benefits, harms).

2. ALL the relevant terms should be defined in an extensive glossary, and the definitions should be both technically accurate and easily understood. The glossary was compiled by Jeremy Howick with help from Hazel Thornton, Ian Chalmers, and two research assistants (Morwenna James, and Katherine Law).

3. That screening tests were sufficiently important to merit a separate entry, and that the importance of systematic reviews should be emphasized. That we should consider all relevant evidence is a fundamental tenet of the scientific method (reproducibility).

After the meeting, Jeremy Howick, Paul Glasziou, and Carl Heneghan drafted a Levels and in January Jeremy Howick sent it to the Working Group for feedback. In March and May 2011 Jeremy Howick posted it on the OCEBM website, and invited subscribers to the CEBM mailing list to comment before September 1st. Jeremy Howick also sent the documents to Gordon Guyatt, Brian Haynes, and Dave Sackett. Brian Haynes made some useful suggestions. On September 1st, Jeremy Howick collated the feedback, made some changes to the Levels, and circulated both the feedback and the revised Levels to the OCEBM Evidence Working Group.

Major Changes to the 2011 OCEBM Levels

What is the same (the good things we didn't change)

1. The rows and columns are switched.

a. Each **row** represents a series of steps to should follow when searching for likely best evidence. The likely strongest evidence is likely to be found furthest to the left of the Levels, and each column to the right represents likely weaker evidence.

b. Each **column** represents the types of questions the clinician is likely to encounter *in the order the clinician will encounter them*. For example, the first question a clinician might want to ask is the prevalence (How common is it?). Then, they might like to know whether the diagnostic test

This must be read before using the Levels: no evidence ranking system or decision tool can be used without a healthy dose of judgment and thought.

@CEBMOXFORD

The @EBMDataLab are looking for a researcher to join their team. Further info: https://t.co/LhajNRqKf8 @OxPrimaryCare @bengoldacre about 1 hour ago

Follow @cebmoxford 3,512 followers

was accurate. Next, they should wonder what would happen if they did not prescribe a therapy, and whether the likely benefits of the treatment they propose outweigh the likely harms.

2. Although busy clinicians might have to resort to individual studies, the OCEBM Levels is NOT dismissive of systematic reviews. On the contrary, systematic reviews are better at assessing strength of evidence than single studies(17, 18) and should be used if available. On the other hand clinicians or patients might have to resort to individual studies if a systematic review is unavailable. GRADE, for example, assumes that there is a systematic review and is of limited use when systematic reviews have not been conducted. The one exception to using a systematic review first is for questions of local prevalence, where a current local survey is ideal.

3. We added questions about common and rare harms, and the value of (early) screening because we felt that these were important and clinically relevant questions.

4. We omitted questions about economic and decision analysis. Although analyses are essential, we felt that further research, perhaps together with economists and policy makers, was required before pronouncing on what counts as good evidence in these areas.

5. We omitted most of the footnotes from the original Levels.

6. A new OCEBM Glossary will accompany the Levels. The new Glossary is more extensive and friendly.

7. We divided harms into 'common' and 'rare'. A rule of thumb is that a common harm involves more than 20% of participants.

Justification for the 2011 OCEBM Levels

Although the 2011 OCEBM Levels is based on what type of evidence is likely to provide strongest support from both empirical (19-21)and theoretical (11, 22, 23) work. In a word, the lower the risk of confounding (bias), the further to the left the type of evidence will lie.

Empirical investigation of OCEBM Levels

While it is difficult to assess the number of citations to the OCEBM Levels because the original document did not provide instructions for how to cite the Levels, a Google search of "Oxford CEBM Levels" yields over 10 000 results, a Google search of "OCEBM Levels" yields over 300 results, and a PubMed search of "Oxford Levels of Evidence" yields 794 results. Systematic reviewers (24-28), clinicians, and policy makers (29) have all used the OCEBM Levels to judge the strength of evidence. Instruction for citing the revised OCEBM Levels is clearer which should make tracking its use more straightforward.

Potential limitations of the 2011 OCEBM Levels

While relatively simple rules of evidence can be more reliable than more complex strategies (15, 16, 30), they are not foolproof. Certainly one can always imagine scenarios where evidence from a column further to the right – say observational studies with dramatic effects – will provide stronger evidence than something currently ranked further to the left – say a systematic review of randomized trials. For example, imagine a systematic review with that didn't include all the relevant studies and was conducted by a potentially biased organization (31) suggested that a treatment had a positive benefit. We might stop our search in the belief that we had found sufficiently strong evidence to make a decision. However, if we continued, we might have found a recent, large, well-conducted randomized trial indicating that the same treatment had no benefit or perhaps was harmful. Which evidence do we accept?

There are two potential answers to this question. One would be to make the Levels more complex by introducing more columns. Instead of having 5 columns, we could have 10, 20, or more. In the different columns we could differentiate all the different 'qualities' of, say, systematic reviews of treatment benefits. For example, we might place systematic reviews of low quality randomized trials in a column to the right (likely worse evidence) than a large, high-quality randomized trial. The problem with introducing more columns is that the Levels would no longer be simple, and clinicians would not be able to usable in real time. Moreover exceptions would never altogether

disappear and empirical investigations might reveal that the simpler hierarchies lead to better average decisions than more complex alternatives.

The other solution is to insist that the Levels be interpreted with a healthy dose of common sense and good judgment (13, 14, 32, 33), which brings us to the next section.

The role of expertise in using the OCEBM Levels

A problem with all hierarchies of evidence is that, psychologically and sociologically speaking, they encourage people to stop using judgment. No hierarchy or levels of evidence can be used without careful thought (34).

This does not imply, of course, that experts should ignore evidence. Indeed both in the original (35) and revised 'Bradford Hill Guidelines' (33), researchers are asked to consider various factors when making clinical decisions. At the same time we believe that a healthy dose of scepticism and judgement will always be required to appraise evidence and apply it to individuals in routine practice (11, 36, 37).

Future directions

The strength of evidence is related to what the evidence is for(11), and good evidence for clinical decisions should answer clinically relevant questions. What the clinician, patient, or policy maker wants to know (amongst other things) is, 'Which treatment, from among all the available alternatives, has the most favourable benefit/harm balance?' For example, surgery may well be effective for back pain, but so may other, less risky treatments (38-40). Or consider depression. There are several selective serotonin reuptake inhibitors (SSRIs) and numerous other pharmacological antidepressants (tricyclics, monoamine oxidase inhibitors (MAOis), serotonin-norepinephrine reuptake inhibitors (SNRIs), noradrenergic and specific serotonergic antidepressants (NASSAs), norepinephrine (noradrenaline) reuptake inhibitors (NRIs), and Norepinephrinedopamine reuptake inhibitors). Then, there are many non-pharmaceutical treatments used to treat depression, including St. John's wort, Cognitive Behaviour Therapy (CBT), exercise, and self-help. None of these treatments has demonstrated consistent superiority to others in trials (41). In order to rationally choose which therapy to use we must understand the relative benefits and harms of these different options.

With this in mind, the row in the Levels about therapeutic benefits should, ideally, be 'Which treatment, from among all available alternatives, has the most favourable benefit/harm balance?' The evidence to answer such a question would most probably come in Levels that included the various treatment options, together with the quality of evidence for benefits and harms. We chose not to include such a row because such evidence is, at the time of writing, rare. Fortunately, evidence comparing all available alternatives is becoming more common in the form of 'umbrella reviews' (42) and 'comparative effectiveness research' (43). We expect that the next version of the OCEBM Levels will ask clinicians to consider the relative benefits and harms of all available alternatives.

Conclusion

The 2011 OCEBM Levels was developed by an international group and took into account feedback from clinicians, patients, and all those on the OCEBM mailing list. It retains the spirit of the original 1998 OCEBM Levels in that it covers a range of clinical questions, it can be used to find the likely best evidence quickly, and it encourages clinicians and patients to assess evidence autonomously. The 6 main changes include reversal of the rows and columns, additional Levels for harms and screening tests, increased simplicity and an extensive glossary.

How to cite the Background Document

Jeremy Howick, Iain Chalmers, Paul Glasziou, Trish Greenhalgh, Carl Heneghan, Alessandro Liberati, Ivan Moschetti, Bob Phillips, and Hazel Thornton. "Explanation of the 2011 Oxford Centre for Evidence-Based Medicine (OCEBM) Levels of Evidence (Background Document)". Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653

References

1. TRIP. The Trip Database. Newport, UK2009 [cited 2009 12 November 2009]; Available from: www.tripdatabase.com.

2. NIH. PubMed. Betheseda: U.S. National Library of Medicine; 2009 [cited 2009 12 November 2009]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/.

3. REHAB+. REHAB+. Hamilton: McMaster University; 2009 [cited 2009 12 November 2009]; Available from: http://plus.mcmaster.ca/Rehab/Default.aspx.

4. Canadian Task Force on the Periodic Health Examination. The periodic health examination. Can Med Assoc J. 1979;121:1193-254.

5. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1986 Feb;89(2 Suppl):2S-3S.

6. Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1989 Feb;95(2 Suppl):2S-4S.

7. Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombotic agents. Chest. 1992 Oct;102(4 Suppl):305S-11S.

8. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008 Apr 26;336(7650):924-6.

9. The periodic health examination. Canadian Task Force on the Periodic Health Examination. Can Med Assoc J. 1979 Nov 3;121(9):1193-254.

10. Harbour RT, editor. SIGN 50: A guideline developer's handbook. Edinburgh: NHS Quality Improvement Scotland; 2008.

11. Howick J. The Philosophy of Evidence-Based Medicine. Oxford: Wiley-Blackwell; 2011.

12. Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ. 2003 Dec 20;327(7429):1459-61.

13. Glasziou P, Chalmers I, Rawlins M, McCulloch P. When are randomised trials unnecessary? Picking signal from noise. BMJ. 2007 Feb 17;334(7589):349-51.

14. Aronson JK, Hauben M. Anecdotes that provide definitive evidence. BMJ. 2006 Dec 16;333(7581):1267-9.

15. Gigerenzer G. Gut feelings : short cuts to better decision making. London: Penguin, 2008; 2007.

16. Gigerenzer G, Todd PM. Simple heuristics that make us smart. New York: Oxford University Press; 1999.

17. Chalmers I. The lethal consequences of failing to make full use of all relevant evidence about the effects of medical treatments: the importance of systematic reviews. In: Rothwell PM, editor. Treating individuals: from randomised trials to personalized medicine. London: The Lancet; 2007.

18. Lane S, Deeks J, Chalmers I, Higgins JP, Ross N, Thornton H. Systematic Reviews. In: Science SA, editor. London 2001.

19. Khan KS, Daya S, Collins JA, Walter SD. Empirical evidence of bias in infertility research: overestimation of treatment effect in crossover trials using pregnancy as the outcome measure. Fertil Steril. 1996 May;65(5):939-45.

20. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995 Feb 1;273(5):408-12.

21. Wood L, Egger M, Gluud LL, Schulz KF, Juni P, Altman DG, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008 Mar 15;336(7644):601-5.

22. La Caze A. Evidence-Based Medicine Must Be. J Med Philos. 2009 Aug 18.

22. La Caze A. Evidence-Based Medicine Must Be. J Med Philos. 2009 Aug 18.

24. Beaton R, Pagdin-Friesen W, Robertson C, Vigar C, Watson H, Harris SR. Effects of exercise intervention on persons with metastatic cancer: a systematic review. Physiother Can. 2009 Summer;61(3):141-53.

25. Moreno L, BauKsta F, Ashley S, Duncan C, Zacharoulis S. Does chemotherapy affect the visual outcome in children with optic pathway glioma? A systematic review of the evidence. Eur J Cancer. Aug;46(12):2253-9.

26. Galderisi S, Mucci A, Volpe U, Boutros N. Evidence-based medicine and electrophysiology in schizophrenia. Clin EEG Neurosci. 2009 Apr;40(2):62-77.

27. Cooper C, Balamurali TB, Livingston G. A systematic review of the prevalence and covariates of anxiety in caregivers of people with dementia. Int Psychogeriatr. 2007 Apr;19(2):175-95.

28. Freeman BJ. IDET: a critical appraisal of the evidence. Eur Spine J. 2006 Aug;15 Suppl 3:S448-57.

29. Naver L, Bohlin AB, Albert J, Flamholc L, Gisslen M, Gyllensten K, et al. Prophylaxis and treatment of HIV-1 infection in pregnancy: Swedish Recommendations 2007. Scand J Infect Dis. 2008;40(6-7):451-61.

30. Gigerenzer G. Reckoning with risk : learning to live with uncertainty. London: Penguin, 2003; 2002.

31. Jorgensen AW, Hilden J, Gotzsche PC. Cochrane reviews compared with industry supported meta-analyses and other meta-analyses of the same drugs: systematic review. BMJ. 2006 Oct 14;333(7572):782.

32. Glasziou P, Vandenbroucke JP, Chalmers I. Assessing the quality of research. BMJ. 2004 Jan 3;328(7430):39-41.

33. Howick J, Glasziou P, Aronson JK. The evolution of evidence hierarchies: what can Bradford Hill's 'guidelines for causation' contribute? J R Soc Med. 2009 May;102(5): 186-94.

34. Hill ABS, Hill ID. Bradford Hill's principles of medical statistics. 12th ed. ed: Edward Arnold; 1991.

35. Hill AB. The Environment and Disease: Association or Causation? Proceedings of the Royal Society of Medicine. 1965;58:295-300.

36. Haynes RB, Devereaux PJ, Guyatt GH. Clinical expertise in the era of evidence-based medicine and patient choice. ACP J Club. 2002 Mar-Apr;136(2):A11-4.

37. Straus SE, Richardson WS, Glasziou P, Haynes RB. Evidence-Based Medicine: How to Practice and Teach EBM. 3rd ed. London: Elsevier: Churchill Livingstone; 2005.

38. Mirza SK, Deyo RA. Systematic review of randomized trials comparing lumbar fusion surgery to nonoperative care for treatment of chronic back pain. Spine (Phila Pa 1976). 2007 Apr 1;32(7):816-23.

39. Brox JI, Reikeras O, Nygaard O, Sorensen R, Indahl A, Holm I, et al. Lumbar instrumented fusion compared with cognitive intervention and exercises in patients with chronic back pain aper previous surgery for disc herniation: a prospective randomized controlled study. Pain. 2006 May;122(1-2):145-55.

40. Fritzell P, Hagg O, Nordwall A. Complications in lumbar fusion surgery for chronic low back pain: comparison of three surgical techniques used in a prospective randomized study. A report from the Swedish Lumbar Spine Study Group. Eur Spine J. 2003 Apr; 12(2):178-89.

41. Howick J. Questioning the Methodologic Superiority of 'Placebo' Over 'Active' Controlled Trials American Journal of Bioethics. 2009;9(9):34-48.

42. Becker L. The Cochrane Colloquium: Umbrella Reviews: What are they, and do we need them? : The Cochrane Collaboration; 2010 [cited 2010 10 September 2010]; Available from: http://www.slideshare.ne /Cochrane.Collaboration/umbrellareviews-what-are-they-and-do-we-need-them-160605.

43. Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. Ann Intern Med. 2009 Aug 4;151(3):203-5.

Downloads

- Introduction to the CEBM-Levels of Evidence (PDF)
- French translations of the Introduction to the CEBM Levels of Evidence

Related Posts

No related posts.

EXHIBIT 4

Essay

The GRADE approach and Bradford Hill's criteria for causation

Holger Schünemann,[1] Suzanne Hill,[2] Gordon Guyatt,[1] Elie A Akl,[3] Faruque Ahmed[4]

[1]Departments of Clinical Epidemiology & Biostatistics and of Medicine, McMaster University Health Sciences Centre, Hamilton, Ontario, Canada
[2]World Health Organization, Geneva, Switzerland
[3]State University of New York, Buffalo, New York, USA
[4]Centers for Disease Control and Prevention, Atlanta, Georgia, USA

Correspondence to
Holger J Schünemann, Department of Clinical Epidemiology & Biostatistics, McMaster University Health Sciences Centre, Room 2C10B, 1200 Main Street West, Hamilton, ON L8N 3Z5, Canada; schuneh@mcmaster.ca

Accepted 19 July 2010
Published Online First 14 October 2010

ABSTRACT
This article describes how the Grading of Recommendations Assessment, Development and Evaluation (GRADE) approach to grading the quality of evidence and strength of recommendations considers the Bradford Hill criteria for causation and how GRADE may relate to questions in public health. A primary concern in public health is that evidence from non-randomised studies may provide a more adequate or best available measure of a public health strategy's impact, but that such evidence might be graded as lower quality in the GRADE framework. GRADE, however, presents a framework that describes both criteria for assessing the quality of research evidence and the strength of recommendations that includes considerations arising from the Bradford Hill criteria. GRADE places emphasis on recommendations and in assessing quality of evidence; GRADE notes that randomisation is only one of many relevant factors. This article describes how causation may relate to developing recommendations and how the Bradford Hill criteria are considered in GRADE, using examples from the public health literature with a focus on immunisation.

> "Scientists believe in proof without certainty; most people believe in certainty without proof." (Ashely Montagu; from http://meds.queensu.ca/medicine/obgyn/links/criteria_for_causation.htm)

It has been proposed that the Grading of Recommendations Assessment, Development and Evaluation (GRADE) for public health questions should consider the Bradford Hill criteria for causation and that GRADE requires adaptation.[1] In this article, we describe the relation of the Bradford Hill criteria to the GRADE approach to grading the quality of evidence and strength of recommendations. The primary concern seems that evidence from non-randomised studies may provide a more adequate or best available measure of a public health strategy's impact, but that such evidence might be graded as lower quality in the GRADE framework. We would like to reiterate that GRADE presents a framework that describes both criteria for assessing the quality of research evidence and the strength of recommendations. In assessing quality of evidence, GRADE notes that randomisation is only one of many relevant factors. Furthermore, GRADE is not specific to the narrow field of therapeutic interventions. Indeed, it likely is the most broadly applied framework for evaluation of evidence and developing recommendations.

We would like to clarify several issues that have been raised in a recent editorial published in the Journal of Epidemiology and Community Health.[1] First, concern has been expressed that herd immunity as a result of immunisation and indirect effects on the co-circulation of other pathogens are typically ascertained through the use of observational epidemiological methods. Although we do not disagree with this assessment, we would like to point out that, innovative randomised controlled trials (RCTs) using cluster-randomisation can be conducted to provide such information.[2] Second, concern is expressed that a quasi-RCT that found a 94% protective effect of a live, monovalent vaccine against measles was classified as 'moderate level of scientific evidence'. However, GRADE's strength of association criteria can be applied to quasi-RCTS and observational studies with no major threats to validity to upgrade the quality of evidence (see below). Such a judgement would be possible in this situation. Third, it is implied that GRADE ratings do not give credit to the 'gradient of effects with scale of population level impact compatible with degree of coverage'. However, we would like to clarify that GRADE's dose-response criterion is not limited to clinical dose only, and that it can be applied to such gradients at the population level to upgrade the quality of the evidence.

Finally, it has been speculated that anti-vaccination lobby groups may abuse the GRADE ratings.[1] Although, abuse of any system is possible, in the case of GRADE it is equally likely that increased transparency provided by the GRADE framework can strengthen, rather than undermine, the trust in vaccines and other clinical interventions.[3 4]

How does GRADE see the relative role of observational studies and RCTs in judging quality of evidence? The GRADE framework applies to all study designs, but randomisation as a methodological approach to protect against bias and confounding is considered very important. Nevertheless, in certain situations observational studies may provide more relevant information than RCTs (eg, situations of long term follow-up and when RCTs only provide very indirect data).

Does a judgement of moderate or low quality evidence preclude such evidence driving a recommendation in favour, or against, an intervention? In the GRADE approach, it does not. It is possible that post-licensure observation studies on long-term or rare serious adverse effects of a vaccine may receive lower evidence grades, but the possibility of a potentially serious harm may be judged sufficient by a guideline panel to make a recommendation against an intervention. In fact, when moving from evidence to recommendations, GRADE does not focus on research evidence only and the framework does not preclude action based on lower evidence

levels. GRADE also acknowledges the wide range of possible judgements a guideline panel may make, while application of the GRADE approach enhances transparency concerning the evidence that was considered and transparency in how judgements regarding the quality of evidence were made.

It has been requested that the Bradford Hill criteria for assessing causality be considered in the GRADE framework.[1] We agree that Bradford Hill's criteria remain, half a century after their description, relevant factors that influence our confidence in a causal relation. Some of these criteria influence our confidence that estimates of effects are correct or that the association between an exposure and an outcome are trustworthy before making a judgement about causality. We note, however, that there are steps between establishing or surmising causality and moving to interventions that act on the perceived causal relation. Establishing a causal relation between an exposure and an outcome does not always allow a confident inference that all methods of modification or removal of the exposure lead to changes in outcomes. This is in particular true for complex interventions that go beyond simple drug interventions. The judgements involved include the confidence that removal of the exposure can be achieved and are included in the judgements about directness in GRADE.

Nevertheless, GRADE has adopted most of Bradford Hill's criteria, some implicitly, others explicitly. However, we realise that the way GRADE incorporates the criteria for causation may not be evident to everyone. We will therefore describe how the Bradford Hill criteria are considered in the GRADE approach and, when they are not, provide a rationale for not considering the particular criterion.

GRADE defines the quality of evidence as the confidence in an estimate of effect (causal relation) from a body of evidence. We will, therefore, use the term upgrading when this confidence is increased and the term downgrading when this confidence is lowered (table 1).

(1) *Strength of the association*. Bradford Hill suggests that a strong association supports causality. This criterion is directly considered in GRADE through upgrading. In the GRADE system, strong associations between an intervention or exposure and an outcome can lead to upgrading the quality of evidence, ie, increases our confidence that the intervention causes a change in the incidence of that outcome. A second criterion, imprecision, which limits our confidence in an effect, even if strong associations are present, is indirectly related to this item in that it lowers our confidence in an association if the magnitude of the effect is uncertain enough to undermine our confidence.

(2) *Consistency*. Bradford Hill suggests that causation is more likely if the results from various research studies are consistent. This criterion is directly considered in GRADE. The GRADE approach suggests downgrading the quality of evidence when there is inconsistent evidence, ie, when studies of similar quality show unexplained heterogeneity in the estimates of effect.

(3) *Temporality or study design suitability*. Bradford Hill describes that there must be a temporal relation between the exposure and outcome. This criterion, usually better than observational studies, in particular if they are not well designed and conducted is indirectly considered in GRADE in at least three ways. First, evidence from randomised controlled trials—which by default establish this temporal relationship—start as higher quality than evidence from studies that do not establish this relationship in GRADE. Second, longitudinal observational studies that include concurrent control groups would likely provide higher quality evidence than cross-sectional studies. Third, GRADE requires the critical consideration of confounders and covariates that may be responsible for a spurious relation when evaluating observational study designs.[3]

(4) *Biological gradient*. As described by Bradford Hill, a biological gradient between an exposure and the magnitude of an effect increases the confidence in causality. GRADE's criterion of upgrading the quality of evidence for a dose–response relationship is a direct application of this principle.

(5) *Specificity*. According to Bradford Hill, causation is more likely if there is a specific outcome related to a specific exposure in that altering the cause alters the disease outcome. In GRADE, this criterion is indirectly considered in the evaluation of whether both the exposure and the outcome were measured directly and by formulating the question and selecting the population, intervention, comparator and outcome in the first place. However, single exposures or interventions are almost invariably related to many outcomes and vice versa. This criterion is not an important criterion for an evaluation of the effects of interventions.

(6) *Biological plausibility*. Whether the association is plausible or not influences causality in the Bradford Hill approach. GRADE does not consider the issue of plausibility in the strict sense as it was included by Bradford Hill. This is, in part, related to the fact that every relation can be described as plausible given that researchers will always think of an explanation once an association is observed. However, GRADE partially considers plausibility in the evaluation of how direct the intervention is related to a surrogate outcome. For instance, we would frequently accept surrogates that have repeatedly responded to interventions in the same way as patient important outcomes. For example, we accept the use of CD4 levels and HIV viral load as acceptable surrogates for mortality and other patient important outcomes, and one of the reasons for this acceptance is the biological plausibility that CD4 levels and HIV viral load are determinants of disease and therapy success. In addition, GRADE considers biological plausibility as a criterion for the evaluation of the believability of an observed subgroup effect. Furthermore, by asking the question of interest and identifying evidence for or against it, the item of biological plausibility is considered indirectly.

(7) *Coherence*. According to Bradford Hill, causation is more likely if what is observed is supported by and in agreement with the natural history of the disease. GRADE does not consider this criterion explicitly but assessing the validity of surrogate outcomes includes these considerations implicitly as well as formulating appropriate healthcare questions. Furthermore, greater emphasis is placed on direct (eg, long-term population

Table 1 Bradford Hill criteria of causality and their relation to the Grading of Recommendations Assessment, Development and Evaluation (GRADE) criteria for upgrading and downgrading

Bradford Hill criteria	Consideration in GRADE
Strength	Strength of association and imprecision in effect estimate
Consistency	Consistency across studies, ie, across different situations (different researchers)
Temporality	Study design, specific study limitations; RCTs fulfil this criterion better than observational studies, properly designed and conducted observational studies
Biological gradient	Dose–response gradient
Specificity	Indirectness
Biological plausibility	Indirectness
Coherence	Indirectness
Experiment	Study design, randomisation, properly designed and conducted observational studies
Analogy	Existing association for critical outcomes will lead to not downgrading the quality, indirectness

Table 2 Interpretation of the Grading of Recommendations Assessment, Development and Evaluation (GRADE)

Interpretation of strong and conditional (weak) recommendations	Strong recommendation	Conditional (weak) recommendation*
For patients	Most individuals in this situation would want the recommended course of action and only a small proportion would not.	The majority of individuals in this situation would want the suggested course of action, but many would not.
For clinicians	Most individuals should receive the intervention. Formal decision aids are not likely to be needed to help individuals make decisions consistent with their values and preferences.	Recognise that different choices will be appropriate for individual patients and that clinicians must help each patient arrive at a management decision consistent with his or her values and preferences. Decision aids may be useful in helping individuals making decisions consistent with their values and preferences.
For policy makers and developers of quality measure	The recommendation can be adapted as policy in most situations. Adherence to this recommendation according to the guideline could be used as a quality criterion or performance indicator.	Policy making will require substantial debate and involvement of various stakeholders. An appropriately documented decision making process could be used as quality indicator.
Interpretation of the categories of the quality of evidence		
High: ⊕⊕⊕⊕	There is high confidence that the true effect lies close to that of the estimate of the effect.	
Moderate: ⊕⊕⊕○	There is moderate confidence in the effect estimate: the true effect is likely to be close to the estimate of the effect, but there is a possibility that it is substantially different.	
Low: ⊕⊕○○	The panel's confidence in the effect estimate is limited: the true effect may be substantially different from the estimate of the effect.	
Very low: ⊕○○○	There is little confidence in the effect estimate: the true effect is likely to be substantially different from the estimate of effect.	

*Guideline panels applying GRADE use the terms 'conditional' and 'weak' synonymously.

important outcomes) rather than short-term outcomes during the formulating of questions and the evaluation of the evidence.
(8) *Experimental evidence*. Experimental evidence enhances the probability of causation. GRADE places emphasis on rigorous experimental designs and this criterion is directly considered. RCTs provide the ideal experimental study design to establish causation where randomisation is the leading experimental factor that increases confidence in associations. Flaws in the experimental design or implementation of an RCT lead to downgrading the quality of evidence. Better experimentally designed observational studies with independent control groups will be graded as higher quality than poorly designed observational studies.
(9) *Reasoning by analogy*. Bradford Hill suggests that existing similar associations would support causation. This criterion is indirectly considered in the GRADE approach. The overall quality of evidence may not be lowered for a single critical outcome if higher quality evidence exists for other critical outcomes and the association is similar in direction.[5] For example, if an intervention to reduce exposure (eg, air pollution) is associated with mortality and chronic respiratory disease and this is based on moderate (⊕⊕⊕○) quality evidence, but only low (⊕⊕○○) quality exists for a third outcome, such as stroke, but all associations are indicating similar effects, then the overall quality would not be lowered because of the single outcome of low quality even if it is critical. Furthermore, GRADE considers indirect evidence when direct evidence is not available.

We appreciate the opportunity to provide clarification regarding how the GRADE framework applies to public health.[1] The GRADE framework—like other evidence-based systems—is an evolving system and we welcome input and insights from users on the strengths and challenges of applying GRADE to vaccines and other preventive public health programmes. Additional use in the field may improve GRADE, in particular in the field of public health and policy interventions, and will advance the field of guidance development. We have previously discussed some of the advantages and disadvantages of applying one approach across different questions.[6] In regard to the Bradford Hill criteria, we believe that the GRADE approach appropriately includes most of the considerations that Bradford Hill suggested.

Finally, there are two other issues that are relevant to this discussion and require emphasis. We remind users of GRADE that the approach separates the quality of evidence from the strength of recommendations and that for the appropriate emphasis we place the recommendation before the quality rating. No recommendation should come without appropriate interpretation (see table 2 for interpretation of the strength of recommendation and the quality of evidence). Guideline developers can (and sometimes should) make strong recommendations on the basis of low or very low quality evidence, but GRADE demands that these situation should be transparently described.

Furthermore, we understand that *labelling* quality as *low* or *very low* may be a valid concern and that other descriptors such as the symbols we presented above (eg, ⊕⊕⊕○ for moderate) may help overcome reluctance to accept the underlying evidence due to labelling issues. We therefore suggest alternatives such as symbols or letters. Details about the GRADE system are published elsewhere, but in this article we have provided a brief guide for those who are dealing with observational study designs.[7–12] The approach has been used by many groups in the public health and policy sector, including guideline panels at WHO.

What this study adds

The GRADE approach to assessing the quality of evidence and grading the strength of healthcare recommendations has been used widely. This article deals with queries regarding it's applicability to public health questions and how to move from studies of exposure-risk assessment to recommendations when using the GRADE framework with special consideration for the Bradford Hill criteria for causation.

Acknowledgements The authors thank Andrew D Oxman for helpful input.

Competing interests HJS is co-chair of the GRADE working group; he supports the implementation of the GRADE approach worldwide. From non-profit organisations he has accepted honoraria and consulting fees for activities in which his work with GRADE may be relevant. SH is a staff member of the WHO. The authors alone are responsible for the views expressed in this publication and they do not necessarily represent the decisions, policy or views of the WHO or other organisations. The conclusions in this article are those of the authors and do not necessarily represent the official position of the US Centers for Disease Control and Prevention. GG is co-chair of the GRADE working; he supports the implementation of the GRADE

approach worldwide. On behalf of McMaster University, he has accepted honoraria and consulting fees for activities in which his work with GRADE is relevant.

Contributors HJS drafted the first version of the article. FA made substantial contributions to the first draft. All other authors made important additional contributions.

Provenance and peer review Not commissioned; not externally peer reviewed.

REFERENCES

1. **Durrheim DN,** Reingold A. Modifying the GRADE framework could benefit public health. *J Epidemiol Community Health* 2010;**64**:387.
2. **Sur D,** Ochiai RL, Bhattacharya SK, *et al*. A cluster-randomized effectiveness trial of Vi typhoid vaccine in India. *N Engl J Med* 2009;**361**:335–44.
3. **Hebert PC,** Levin AV, Robertson G. Bioethics for clinicians: 23. Disclosure of medical error. *CMAJ* 2001;**164**:509–13.
4. **Lopez L,** Weissman JS, Schneider EC, *et al*. Disclosure of hospital adverse events and its association with patients' ratings of the quality of care. *Arch Intern Med* 2009;**169**:1888–94.
5. **Brozek J,** Oxman A, Schünemann HJ. GRADEpro. [Computer program]. Version 3.2 for Windows. http://mcmaster.flintbox.com/technology.asp?Page=3993 and http://www.cc-ims.net/revman/gradepro (accessed 28 Mar 2011).
6. **Schünemann H,** Fretheim A, Oxman AD. Improving the use of research evidence in guideline development: 9. Grading evidence and recommendations. *Health Res Policy Syst* 2006;**4**:21.
7. **Guyatt GH,** Oxman AD, Kunz R, *et al*. Going from evidence to recommendations. *BMJ* 2008;**336**:1049–51.
8. **Guyatt GH,** Oxman AD, Kunz R, *et al*. Incorporating considerations of resources use into grading recommendations. *Br Med J* 2008;**336**:1170–3.
9. **Guyatt GH,** Oxman AD, Kunz R, *et al*. What is "quality of evidence" and why is it important to clinicians? *Br Med J* 2008;**336**:995–8.
10. **Guyatt GH,** Oxman AD, Vist GE, *et al*. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Br Med J* 2008;**336**:924–6.
11. **Schunemann HJ,** Hill SR, Kakad M, *et al*. Transparent development of the WHO rapid advice guidelines. *PLoS Med* 2007;**4**:e119.
12. **Schunemann HJ,** Oxman AD, Brozek J, *et al*. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. *Br Med J* 2008;**336**:1106–10.